https://www.gla.ac.uk/schools/healthwellbeing/
https://www.linkedin.com/in/ojsig/

Using model-based clustering to identify hidden subgroups in Health Data

Author: Jude Chinedu Onunkwo
Supervisor: Rona J. Strawbridge, PhD

Background

- Individuals with severe mental illness are at increased risk of cardiometabolic diseases such as type 2 diabetes (T2D) and cardiovascular diseases (CVDs), independent of social determinants and lifestyle factors[1]
- In a recent study, Strawbridge *et al*.[2] explored whether genetic variants associated with both psychiatric and cardiometabolic diseases could be used to stratify individuals into metabolically distinct subgroups. Using multidimensional scaling (MDS) on data from a high-CVD-risk cohort (IMPROVE) and a general population cohort (UK Biobank), they identified distinct genetic clusters
- However, MDS has limitations in scalability, interpretability, and clinical deployment – motivating the exploration of alternative, more flexible clustering approaches

Objective

- To apply model-based clustering (R package *mclust)* to identify subgroups within a high-CVD-risk cohort

Why this is important

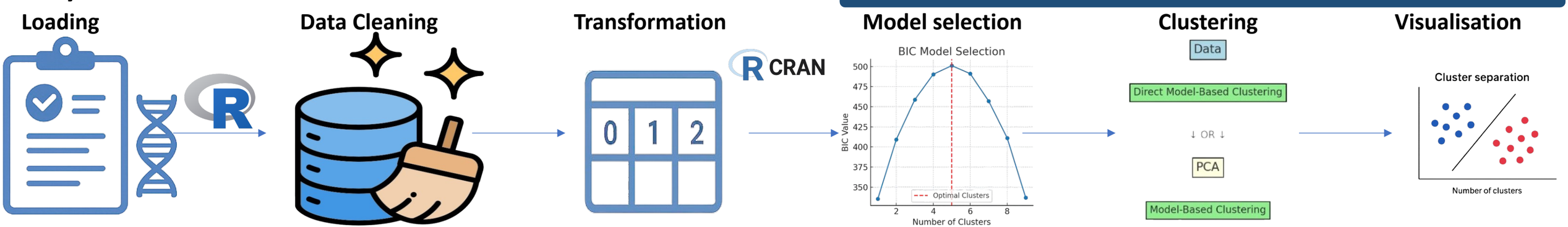
- Genetic variation shapes disease risk and treatment response; clustering reveals hidden patterns in complex single nucleotide polymorphism (SNP) datasets
- Identifying metabolically distinct subgroups in high-CVD/psychiatric-risk populations could support earlier, targeted interventions
- Unlike MDS, model-based clustering provides:
  - Probabilistic group assignments
  - Flexibility in covariance structure
  - Clear statistical criteria (e.g., BIC) for optimal model selection

Methods

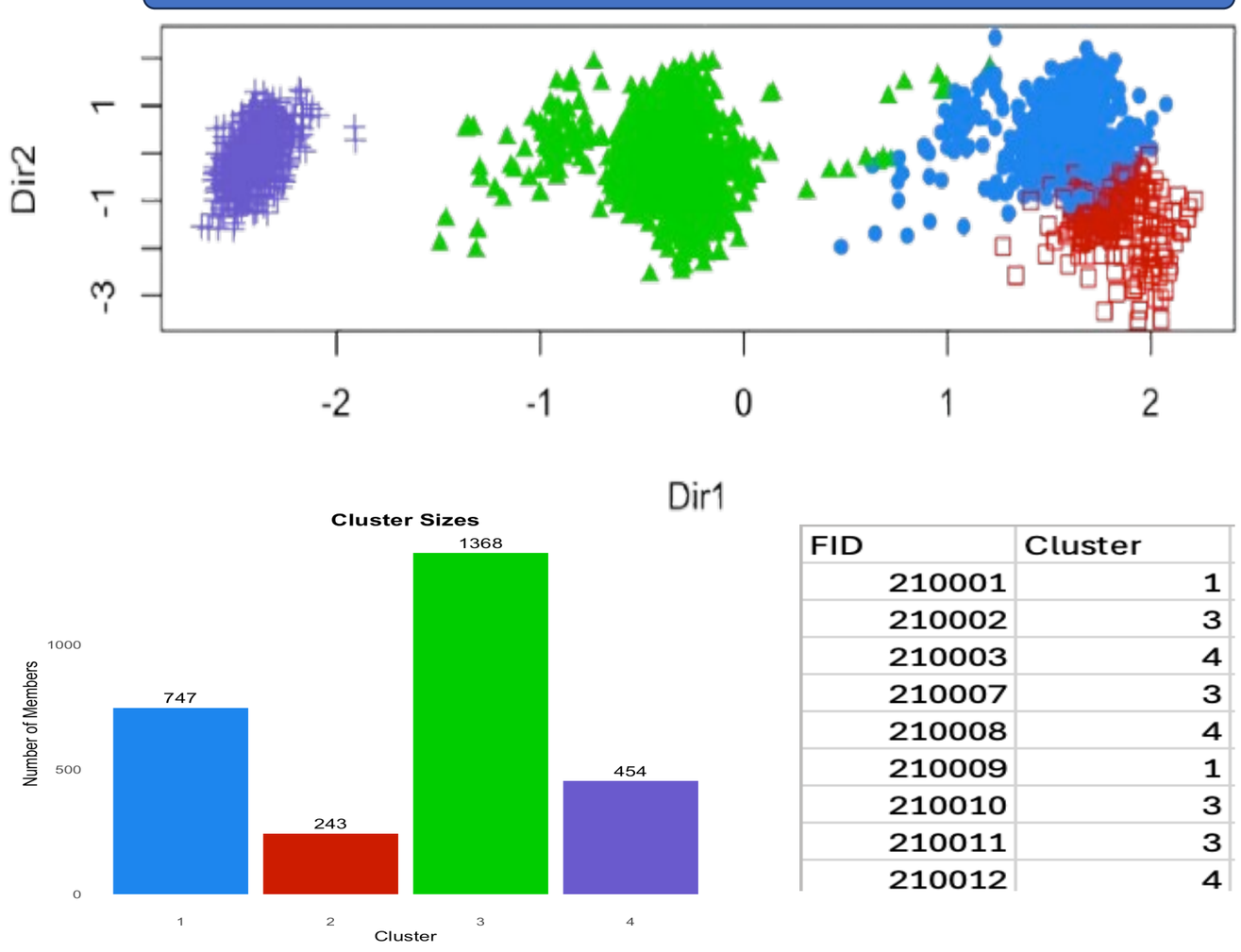
Data Source

- IMPROVE study[3]: 3,700 Europeans (aged 54–79 at high CVD-risk)
- Baseline: Lifestyle/medical questionnaires, anthropometric measures, blood sampling for DNA extraction and clinical biochemistry
- Genotyping: Illumina CardioMetabo + Immuno chips, covering a broad range of cardiometabolic loci
- SNPs: Selected from loci previously linked to severe mental illness and cardiometabolic diseases; strict QC applied

Analysis Workflow



Results and Discussion



FID	Cluster
210001	1
210002	3
210003	4
210007	3
210008	4
210009	1
210010	3
210011	3
210012	4

- Four well-separated sub-groups were identified within the IMPROVE cohort
- The model-based approach is **fast, automated and reproducible,** yielding results consistent with Strawbridge, R *et al*.[2]
- **Limitation:** High dimensional data initially collapsed into one cluster. PCA resolved this but introduced an extra step requiring dimension selection.

Conclusion and Future Directions

- *mclust* is an efficient, fully automated alternative to MDS, offering computational speed and reproducibility
- Findings reinforce shared genetic architecture between mental illness and cardiometabolic diseases, supporting stratified health approaches
- Next steps:
  - Apply to UK Biobank for robustness
  - Characterise subgroup features
  - Explore clinical translation through stratified clinical trials, personalised medicine, and drug development

References:

1. De Hert, M. et al. (2011). Physical illness in patients with severe mental disorders.

2. Strawbridge, R. J. et al. (2021). The overlap of genetic susceptibility to schizophrenia and cardiometabolic disease can be used to identify metabolically different groups of individuals

3. Baldassarre, D. et al. (2010) Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study